BUNDESREPUBLIK DEUTSCHLAND

DE99/1308

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

EJU

Bescheinigung

Die Siemens Aktiengesellschaft in München/Deutschland hat eine Patentanmeldung unter der Bezeichnung

"Verfahren und Anordnung zur Bestimmung spektraler Sprachcharakteristika in einer gesprochenen Äußerung"

am 11. Mai 1998 beim Deutschen Patent- und Markenamt eingereicht.

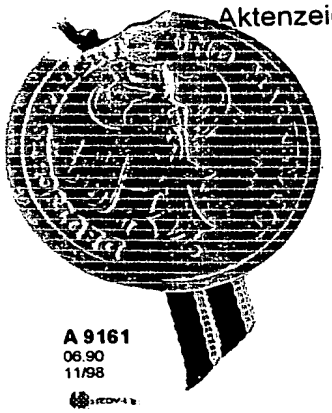
Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

Die Anmeldung hat im Deutschen Patent- und Markenamt vorläufig das Symbol G 10 L 5/00 der Internationalen Patentklassifikation erhalten.

München, den 10. Juni 1999
Deutsches Patent- und Markenamt
Der Präsident
Im Auftrag

Agurks

Aktenzeichen: 198 21 031.0



A 9161
06.90
11/98

THIS PAGE BLANK (USPTO)

Beschreibung

Verfahren und Anordnung zur Bestimmung spektraler Sprachcharakteristika in einer gesprochenen Äußerung

Die Erfindung betrifft ein Verfahren und eine Anordnung zur Bestimmung spektraler Sprachcharakteristika in einer gesprochenen Äußerung.

Bei einer konkatenativen Sprachsynthese werden einzelne Laute aus Sprachdatenbanken zusammengesetzt. Um dabei einen für das menschliche Ohr natürlich klingenden Sprachverlauf zu erhalten, sind Diskontinuitäten an den Punkten, wo die Laute zusammengesetzt werden (Konkatenationspunkte) zu vermeiden. Die Laute sind dabei insbesondere Phoneme einer Sprache oder eine Zusammensetzung mehrerer Phoneme.

Eine Wavelet-Transformation ist aus [1] bekannt. Bei der Wavelet-Transformation ist durch ein Wavelet-Filter gewährleistet, daß jeweils ein Hochpaßanteil und ein Tiefpaßanteil einer nachfolgenden Transformationsstufe ein Signal einer aktuellen Transformationsstufe vollständig wiederherstellen. Dabei erfolgt von einer Transformationsstufe zur nächsten eine Reduktion der Auflösung des Hochpaßanteils bzw. Tiefpaßanteils (engl. Fachbegriff: "Subsampling"). Insbesondere ist durch das Subsampling die Anzahl der Transformationsstufen endlich.

Die **Aufgabe** der Erfindung besteht darin, ein Verfahren und eine Anordnung zur Bestimmung spektraler Sprachcharakteristika anzugeben, mit deren Hilfe insbesondere eine natürlich wirkende synthetische Sprachausgabe bestimmbar ist.

Diese Aufgabe wird gemäß den Merkmalen der unabhängigen Patentansprüche gelöst.

Im Rahmen der Erfindung wird ein Verfahren angegeben zur Bestimmung spektraler Sprachcharakteristika in einer gesprochenen Äußerung. Dazu wird die gesprochenen Äußerung digitalisiert und einer Wavelet-Transformation unterzogen. Anhand unterschiedlicher Transformationsstufen der Wavelet-Transformation werden die sprecherspezifischen Charakteristika ermittelt.

Dabei ist es insbesondere ein Vorteil, daß bei der Wavelet-Transformation mittels eines Hochpaßfilters und eines Tiefpaßfilters die Äußerung aufgeteilt wird und unterschiedliche Hochpaßanteile bzw. Tiefpaßanteile verschiedener Transformationsstufen sprecherspezifische Charakteristika enthalten.

Die einzelnen Hochpaßanteile bzw. Tiefpaßanteile verschiedener Transformationsstufen stehen für vorgegebene sprecherspezifische Charakteristika, wobei sowohl Hochpaßanteil als auch Tiefpaßanteil einer jeweiligen Transformationsstufe, also das jeweilige Charakteristikum, getrennt von anderen Charakteristika modifiziert werden kann. Setzt man bei der inversen Wavelet-Transformation aus den jeweiligen Hochpaß- und Tiefpaßanteilen der einzelnen Transformationsstufen wieder das ursprüngliche Signal zusammen, so ist gewährleistet, daß genau das gewünschte Charakteristikum verändert worden ist. Es ist somit möglich bestimmte vorgegebene Eigenarten der Äußerung zu verändern, ohne daß dadurch der Rest der Äußerung beeinflußt wird.

Eine Ausgestaltung besteht darin, daß vor der Wavelet-Transformation die Äußerung gefenstert, also eine vorgegebene Menge von Abtastwerten ausgeschnitten, und in den Frequenzbereich transformiert wird. Hierzu wird insbesondere eine Fast-Fourier-Transformation (FFT) angewandt.

Eine weitere Ausgestaltung besteht darin, daß ein Hochpaßanteil einer Transformationsstufe in einen Realteil

und einen Imaginärteil aufgeteilt wird. Der Hochpaßanteil der Wavelet-Transformation entspricht dem Differenzsignal zwischen dem aktuellen Tiefpaßanteil und dem Tiefpaßanteil der vorhergehenden Transformationsstufe.

Insbesondere besteht eine Weiterbildung darin, die Zahl der durchzuführenden Transformationsstufen der Wavelet-Transformation dadurch zu bestimmen, daß in der letzten Transformationsstufe, die aus hintereinandergeschalteten Tiefpässen besteht, ein Gleichanteil der Äußerung enthalten ist. Dann ist das Signal als Ganzes darstellbar durch seine Wavelet-Koeffizienten. Dies entspricht der vollständigen Transformation der Information des Signalausschnitts in den Wavelet-Raum.

Wird insbesondere nur der jeweilige Tiefpaßanteil weiter transformiert (mittels eines Hochpaß- und eines Tiefpaßfilters), so verbleibt als Hochpaßanteil einer Transformationsstufe das Differenzsignal, wie oben erläutert. Kumuliert man Differenzsignale (Hochpaßanteile) über die Transformationsstufen, erhält man in der letzten Transformationsstufe als kumulierten Hochpaßanteil die Information der gesprochenen Äußerung ohne Gleichanteil.

Im Rahmen einer zusätzlichen Weiterbildung sind die sprecherspezifischen Charakteristika identifizierbar als:

a) Grundfrequenz:

Die Schwingung des Hochpaßanteils der ersten oder der zweiten Transformationsstufe der Wavelet-Transformation läßt die Grundfrequenz der Äußerung erkennen. Die Grundfrequenz zeigt an, ob der Sprecher ein Mann oder einen Frau ist.

b) Form der spektralen Hüllkurve:

Die spektrale Hüllkurve enthält Information über eine Transferfunktion des Vokaltrakts bei der Artikulation.

In einem stimmhaften Bereich wird die spektrale Hüllkurve von den Formanten dominiert. Der Hochpaßanteil einer höheren Transformationsstufe der Wavelet-Transformation enthält diese spektrale Hüllkurve.

c) Spectral Tilt (Rauchigkeit):
Die Rauchigkeit in einer Stimme wird als negative Steigung im Verlauf des vorletzten Tiefpaßanteils sichtbar.

Die sprecherspezifischen Charakteristika a) bis c) sind bei der Sprachsynthese von großer Bedeutung. Wie eingangs erwähnt, bedient man sich bei der konkatenativen Sprachsynthese großer Mengen realgesprochener Äußerungen, aus denen Beispiellaute ausgeschnitten und später zu einem neuen Wort zusammengesetzt werden (synthetisierte Sprache). Dabei sind Diskontinuitäten zwischen zusammengesetzten Lauten von Nachteil, da diese vom menschlichen Ohr als unnatürlich wahrgenommen werden. Um den Diskontinuitäten entgegenzuwirken ist es von Vorteil, direkt die perzeptiv relevanten Größen zu erfassen und ggf. zu vergleiche und/oder einander anzupassen.

Dies kann geschehen durch direkte Manipulation, indem ein Sprachlaut in mindestens einer seiner sprecherspezifischen Charakteristika angepaßt wird, so daß er in dem akustischen Kontext der konkatenativ verknüpften Laute nicht als störend wahrgenommen wird. Auch ist es möglich, die Auswahl eines passenden Lautes daran auszurichten, daß sprecherspezifische Charakteristika von zu verknüpfenden Lauten möglichst gut zueinander passen, z.B. daß den Lauten gleiche oder ähnliche Rauchigkeit zu eigen ist.

Ein Vorteil der Erfindung besteht darin, daß die spektrale Hüllkurve den Artikulationstrakt des Sprechers widerspiegelt und nicht, wie z.B. ein Polstellenmodell, auf Formanten gestützt ist. Weiterhin gehen bei der Wavelet-Transformation

als nichtparametrischer Darstellung keine Daten verloren, die Äußerung kann stets vollständig rekonstruiert werden. Die aus den einzelnen Transformationsstufen der Wavelet-Transformation hervorgehenden Daten sind linear voneinander unabhängig, können somit getrennt voneinander beeinflußt und später wieder zu der beeinflußten Äußerung - verlustlos - zusammengesetzt werden.

Weiterhin wird eine Anordnung zur Bestimmung spektraler Sprachcharakteristika angegeben, die eine Prozessoreinheit aufweist, die derart eingerichtet ist, daß eine Äußerung digitalisierbar ist. Daraufhin wird die Äußerung einer Wavelet-Transformation unterzogen und anhand unterschiedlicher Transformationsstufen werden sprecherspezifische Charakteristika ermittelt.

Diese Anordnung ist insbesondere geeignet zur Durchführung des erfindungsgemäßen Verfahrens oder einer seiner vorstehend erläuterten Weiterbildungen.

Weiterbildungen der Erfindung ergeben sich auch aus den abhängigen Ansprüchen.

Ausführungsbeispiele der Erfindung werden nachfolgend anhand der Zeichnung dargestellt und erläutert.

Es zeigen

Fig.1 eine Wavelet-Funktion;

Fig.2 eine Wavelet-Funktion, unterteilt nach Realteil und Imaginärteil;

Fig.3 eine kaskadierte Filterstruktur, die die Transformationsschritte der Wavelet-Transformation darstellt;

Fig.4 Tiefpaßanteile und Hochpaßanteile unterschiedlicher Transformationsstufen;

Fig.5 Schritte der konkatenativen Sprachsynthese.

Fig.1 zeigt eine Wavelet-Funktion, die bestimmt ist durch

$$\psi(f) = c \cdot \left(1 - \left(\frac{f}{\sigma}\right)^2\right) \cdot e^{-\frac{1}{2}\cdot\left(\frac{f}{\sigma}\right)^2} \qquad (1),$$

wobei

f die Frequenz,

σ eine Standardabweichung und

c eine vorgegebene Normierungskonstante

bezeichnen.

Insbesondere ist die Standardabweichung σ bestimmt durch die vorgebbare Stelle des Seitenbandminimums 101 in Fig.1.

Fig.2 zeigt eine Wavelet-Funktion mit einem Realteil gemäß Gleichung (1) und einer Hilbert-Transformierten **H** des Realteils als Imaginärteil. Die komplexe Wavelet-Funktion ergibt sich somit zu

$$\Psi(f) = \psi(f) + j \cdot \mathbf{H}\{\psi(f)\} \qquad (2).$$

Die Konstante c aus Gleichung (1) wird verwendet, um die komplexe Wavelet-Funktion zu normieren:

$$\int_{-\infty}^{\infty} \Psi(f) \cdot \overline{\Psi}(f)\, df = 1 \qquad (3),$$

wobei $\overline{\Psi}$ die konjugiert komplexe Wavelet-Funktion bezeichnet.

Fig.3 zeigt die kaskadierte Anwendung der Wavelet-Transformation. Ein Signal 301 wird sowohl durch einen Hochpaß HP1 302 als auch durch einen Tiefpaß TP1 305 gefiltert. Dabei findet insbesondere ein Subsampling statt, d.h. die Anzahl der abzuspeichernden Werte wird pro Filter reduziert. Eine inverse Wavelet-Transformation gewährleistet, daß aus dem Tiefpaßanteil TP1 305 und dem Hochpaßanteil HP1 304 wieder das ursprüngliche Signal 301 rekonstruierbar ist.

Im Hochpaß HP1 302 wird getrennt nach Realteil Re1 303 und Imaginärteil Im1 304 gefiltert.

Das Signal 310 nach dem Tiefpaßfilter TP1 305 wird erneut sowohl durch einen Hochpaß HP2 306 als auch durch einen Tiefpaß TP2 309 gefiltert. Der Hochpaß HP2 306 umfaßt wieder einen Realteil Re2 307 und einen Imaginärteil Im2 308. Das Signal nach der zweiten Transformationsstufe 311 wird wieder gefiltert, usf.

Geht man von einem (FFT-transformierten) Kurzzeitspektrum mit 256 Werten aus, so werden acht Transformationsschritte durchgeführt (Subsamplingrate: 1/2), bis das Signal aus dem letzten Tiefpaßfilter TP8 dem Gleichanteil entspricht.

In **Fig.4** sind verschiedene Transformationsstufen der Wavelet-Transformation, unterteilt nach Tiefpaßanteilen (Figuren 4A, 4C und 4E) und Hochpaßanteilen (Figuren 4B, 4D und 4F) dargestellt.

Aus dem Hochpaßanteil gemäß Fig.4B ist die Grundfrequenz der gesprochenen Äußerung ersichtlich. Neben den Schwankungen in der Amplitude ist deutlich eine überwiegende Periodizität im wavelet-gefilterten Spektrum zu erkennen, die Grundfrequenz des Sprechers. Anhand der Grundfrequenz ist es möglich, vorgegebene Äußerungen bei der Sprachsynthese einander

anzupassen oder passende Äußerungen aus einer Datenbank mit vorgegebene Äußerungen zu bestimmen.

Im Tiefpaßanteil von Fig.4C sind als ausgeprägte Minima und Maxima die Formanten des Sprachsignalausschnitts (die Länge des Sprachsignalausschnitts entspricht in etwa der doppelten Grundfrequenz) dargestellt. Die Formanten repräsentieren Resonanzfrequenzen im Vokaltrakt des Sprechers. Die deutliche Darstellbarkeit der Formanten ermöglicht eine Anpassung und/oder Auswahl passender Lautbausteine bei der konkatenativen Sprachsynthese.

Im Tiefpaßanteil der vorletzten Transformationsstufe (bei 256 Frequenzwerten im Originalsignal: TP7), kann die Rauchigkeit einer Stimme ermittelt werden. Der Abstieg des Kurvenverlaufs zwischen Maximum Mx und Minimum Mi kennzeichnet den Grad der Rauchigkeit.

Die erwähnten drei sprecherspezifischen Charakteristika sind somit identifiziert und können für die Sprachsynthese gezielt beeinflußt werden. Dabei ist es insbesondere von Bedeutung, daß bei der inversen Wavelet-Transformation die Manipulation eines einzelnen sprecherspezifischen Charakteristikums nur dieses beeinflußt, die anderen perziptiv relevanten Größen bleiben unberührt. Somit kann die Grundfrequenz gezielt verstellt werden, ohne daß dadurch die Rauchigkeit der Stimme beeinflußt wird.

Eine andere Einsatzmöglichkeit besteht in der Auswahl eines geeigneten Lautabschnitts zur konkatenativen Verknüpfung mit einem anderen Lautabschnitt, wobei beide Lautabschnitte ursprünglich von verschiedenen Sprechern in unterschiedlichen Kontexten aufgenommen wurden. Mit Ermittlung spektraler Sprachcharakteristika kann ein geeigneter zu verknüpfender Lautabschnitt gefunden werden, da mit den Charakteristika Kriterien bekannt sind, die einen Vergleich von Lautabschnitten untereinander und somit eine Auswahl des

passenden Lautabschnitts automatisch nach bestimmten Vorgaben ermöglichen.

Fig.5 zeigt Schritte einer konkatenativen Sprachsynthese. Eine Datenbank wird mit einer vorgegebenen Menge natürlichgesprochener Sprache verschiedener Sprecher erstellt, wobei Lautabschnitte in der natürlichgesprochenen Sprache identifiziert und abgespeichert werden. Es ergeben sich zahlreiche Repräsentanten für die verschiedenen Lautabschnitte einer Sprache, auf die die Datenbank zugreifen kann. Die Lautabschnitte sind insbesondere Phoneme einer Sprache oder eine Aneinanderreihung solcher Phoneme. Je kleiner der Lautabschnitt, desto größer sind die Möglichkeiten bei der Zusammensetzung neuer Wörter. So umfaßt die deutsche Sprache eine vorgegebene Menge von ca. 40 Phonemen, die zur Synthese nahezu aller Wörter der Sprache ausreichen. Dabei sind unterschiedliche akustische Kontexte zu berücksichtigen, je nachdem, in welchem Wort das jeweilige Phonem auftritt. Nun ist es wichtig, die einzelnen Phoneme in den akustischen Kontext derart einzubetten, daß Diskontinuitäten, die vom menschlichen Gehör als unnatürlich und "synthetisch" empfunden werden, vermieden werden. Wie erwähnt stammen die Lautabschnitte von unterschiedlichen Sprechern und weisen somit verschiedene sprecherspezifische Charakteristika auf. Um eine möglichst natürlich wirkende Äußerung zu synthetisieren, ist es wichtig, die Diskontinuitäten zu minimieren. Dies kann erfolgen durch Anpassung der identifizierbaren und modifizierbaren sprecherspezifischen Charakteristika oder durch Auswahl passender Lautabschnitte aus der Datenbank, wobei ebenfalls die sprecherspezifischen Charakteristika bei der Auswahl ein entscheidendes Hilfsmittel darstellen.

In Fig.5 sind beispielhaft zwei Laute A 507 und B 508 dargestellt, die jeweils einzelne Lautabschnitte 505 bzw. 506 aufweisen. Die Laute A 507 und B 508 stammen jeweils aus einer gesprochenen Äußerung, wobei der Laut A 507 deutlich

vom Laut B 508 verschieden ist. Eine Trennlinie 509 zeigt an, wo der Laut A 507 mit dem Laut B 508 verknüpft werden soll. Im vorliegenden Fall sollen die ersten drei Lautabschnitte des Lautes A 507 mit den letzten drei Lautabschnitten des Lautes B 508 konkatenativ verknüpft werden.

Es wird entlang der Trennlinie 509 ein zeitliches Dehnen oder Stauchen (vergleiche Pfeil 503) der aufeinanderfolgenden Lautabschnitte durchgeführt, um den diskontinuierlichen Eindruck am Übergang 509 zu vermindern.

Eine Variante besteht in einem abrupten Übergang der entlang der Trennlinie 509 geteilten Laute. Dabei kommt es jedoch zu den erwähnten Diskontinuitäten, die das menschliche Gehör als störend wahrnimmt. Fügt man hingegen einen Laut C zusammen, daß die Lautabschnitte innerhalb eines Übergangsbereichs 501 oder 502 berücksichtigt werden, wobei ein spektrales Abstandsmaß zwischen zwei einander zuordenbaren Lautabschnitten in dem jeweiligen Übergangsbereich 501 oder 502 angepaßt wird (allmählicher Übergang zwischen den Lautabschnitten). Als das Abstandsmaß herangezogen wird insbesondere im Wavelet-Raum der euklidische Abstand zwischen den in diesem Bereich relevanten Koeffizienten.

Literaturverzeichnis:

[1] I. Daubechies: "Ten Lectures on Wavelets", Siam Verlag 1992, ISBN 0-89871-274-2, Kapitel 5.1, Seiten 129-137.

Patentansprüche

1. Verfahren zur Bestimmung spektraler Sprachcharakteristika in einer gesprochenen Äußerung,
   a) bei dem die Äußerung digitalisiert wird,
   b) bei dem die digitalisierte Äußerung einer Wavelet-Transformation unterzogen wird,
   c) bei dem anhand unterschiedlicher Transformationsstufen der Wavelet-Transformation die sprecherspezifischen Charakteristika bestimmt werden.

2. Verfahren nach Anspruch 1,
   bei dem vor der Wavelet-Transformation eine gefensterte Transformation der digitalisierten Äußerung in einen Frequenzbereich durchgeführt wird.

3. Verfahren nach Anspruch 2,
   bei dem die Transformation in den Frequenzbereich mittels Fast-Fourier-Transformation durchgeführt wird.

4. Verfahren nach einem der vorhergehenden Ansprüche,
   bei dem in jeder Stufe der Wavelet-Transformation ein Tiefpaßanteil und ein Hochpaßanteil eines zu transformierenden Signals ermittelt werden.

5. Verfahren nach einem der vorhergehenden Ansprüche,
   bei dem ein Hochpaßanteil nach einem Realteil und einem Imaginärteil unterteilt wird.

6. Verfahren nach einem der vorhergehenden Ansprüche,
   bei dem die Wavelet-Transformation mehrere Transformationsstufen umfaßt, wobei die letzte Transformationsstufe einen Gleichanteil der Äußerung in einer der Anzahl Transformationsstufen entsprechenden wiederholten Tiefpaßfilterung liefert.

7. Verfahren nach einem der vorhergehenden Ansprüche, bei dem die sprecherspezifischen Charakteristika bestimmt sind durch:
   a) eine Grundfrequenz der gesprochenen Äußerung;
   b) spektrale Hüllkurve;
   c) einer Rauchigkeit der gesprochenen Äußerung.

8. Verwendung des Verfahrens nach einem der Ansprüche 1 bis 7 zur Sprachsynthese,
   wobei einzelne sprecherspezifische Charakteristika im Hinblick auf eine natürlich klingende Aneinanderreihung von Sprachlauten angepaßt werden.

9. Verwendung des Verfahrens nach einem der Ansprüche 1 bis 7 zur Sprachsynthese,
   wobei aus einer vorgegebenen Datenmenge diejenigen Sprachlaute anhand einzelner spektraler Sprachcharakteristika ausgewählt werden, die eine natürlich klingende Aneinanderreihung von Sprachlauten gewährleisten.

10. Anordnung zur Bestimmung spektraler Sprachcharakteristika in einer gesprochenen Äußerung
    mit einer Prozessoreinheit, die derart eingerichtet ist, daß folgende Schritte durchführbar sind:
    a) die Äußerung wird digitalisiert;
    b) die digitalisierte Äußerung wird einer Wavelet-Transformation unterzogen;
    c) anhand unterschiedlicher Transformationsstufen der Wavelet-Transformation werden die sprecherspezifischen Charakteristika bestimmt.

Zusammenfassung

Verfahren und Anordnung zur Bestimmung spektraler Sprachcharakteristika in einer gesprochenen Äußerung

Es werden spektrale Sprachcharakteristika in einer natürlichsprachlichen Äußerung bestimmt, wobei die Äußerung digitalisiert und einer Wavelet-Transformation unterzogen wird. Aus den unterschiedlichen Transformationsstufen der Wavelet-Transformation gehen die sprecherspezifischen Charakteristika hervor. Diese Charakteristika können im Rahmen einer Sprachsynthese mit Charakteristika anderer Äußerungen verglichen werden, um ein für das menschliche Ohr kontinuierlich klingendes synthetisches Sprachsignal zu erzeugen. Alternativ können die Charakteristika auch gezielt verändert werden, um einer perzeptiven Dissonanz entgegenzuwirken.

FIG 1

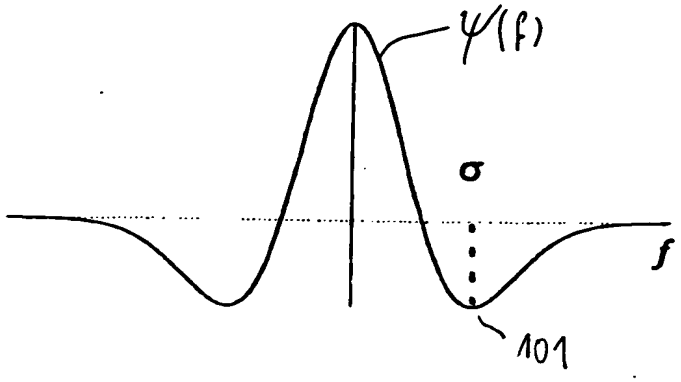


FIG 2

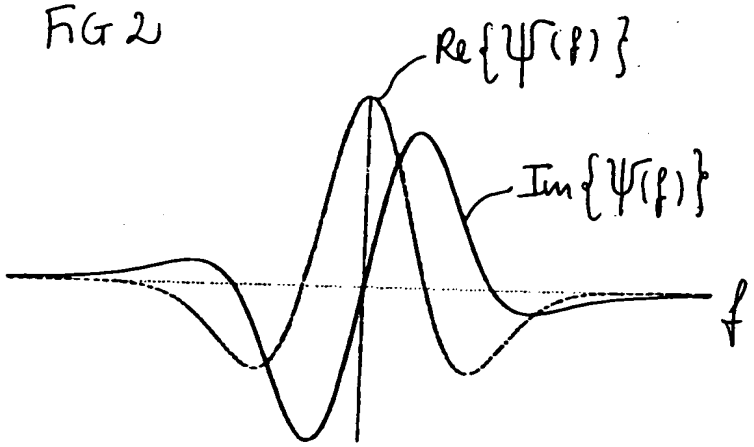
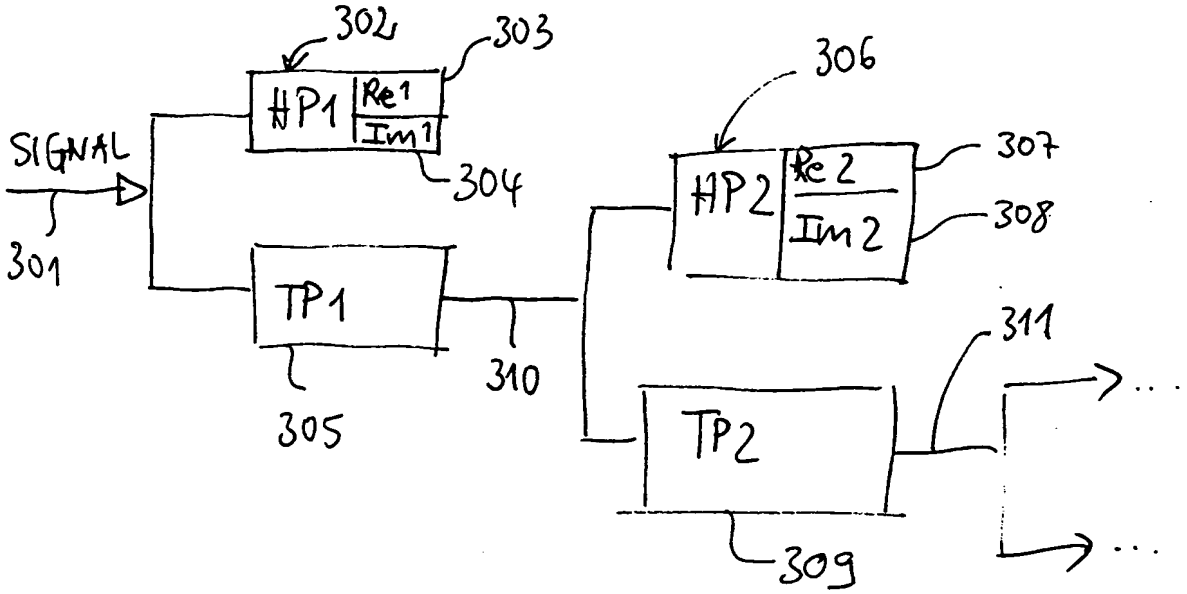


FIG 3


302
303
306
HP1
Re1
Im1
SIGNAL
304
HP2
Re2
Im2
307
308
301
TP1
310
311
305
TP2
309

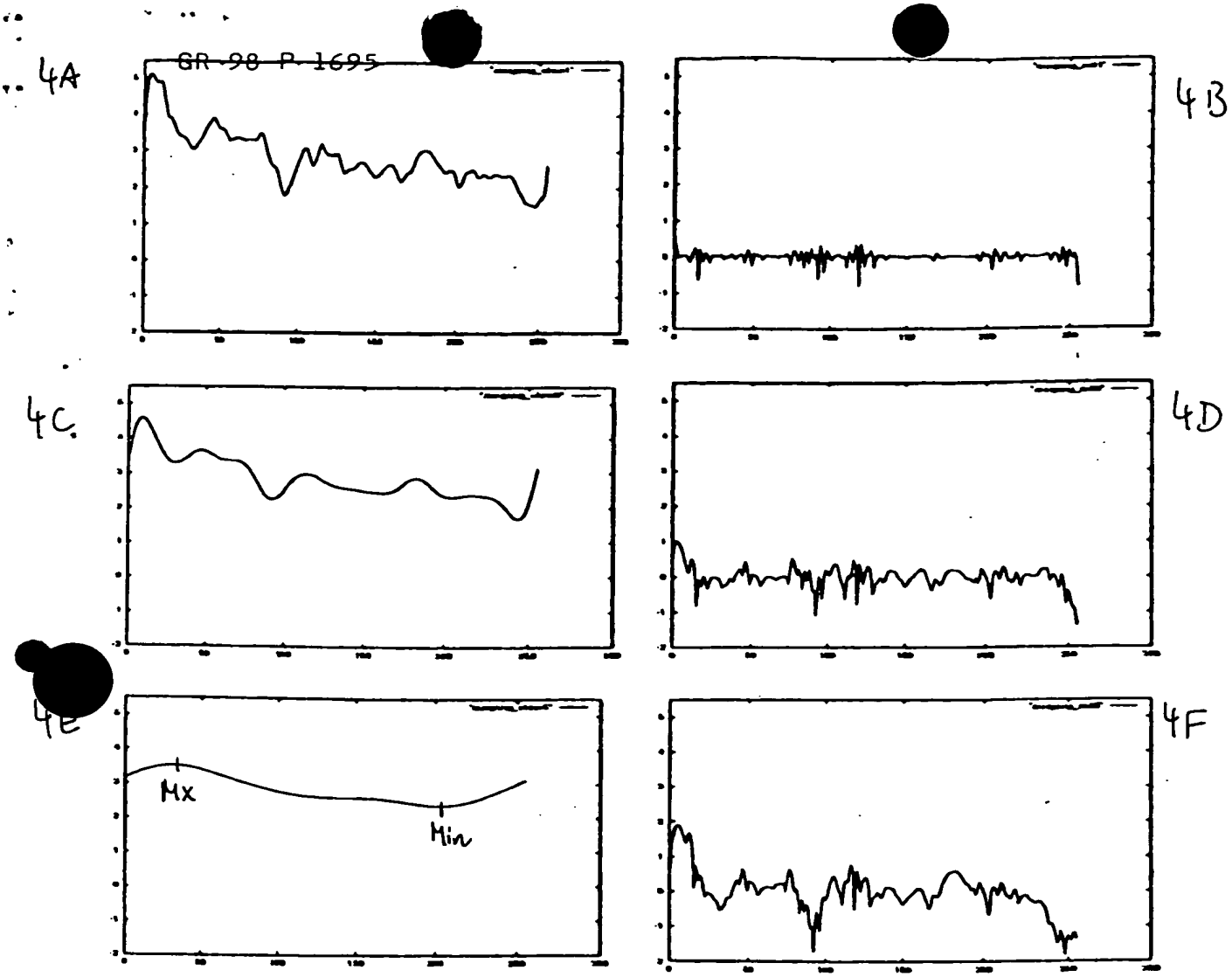
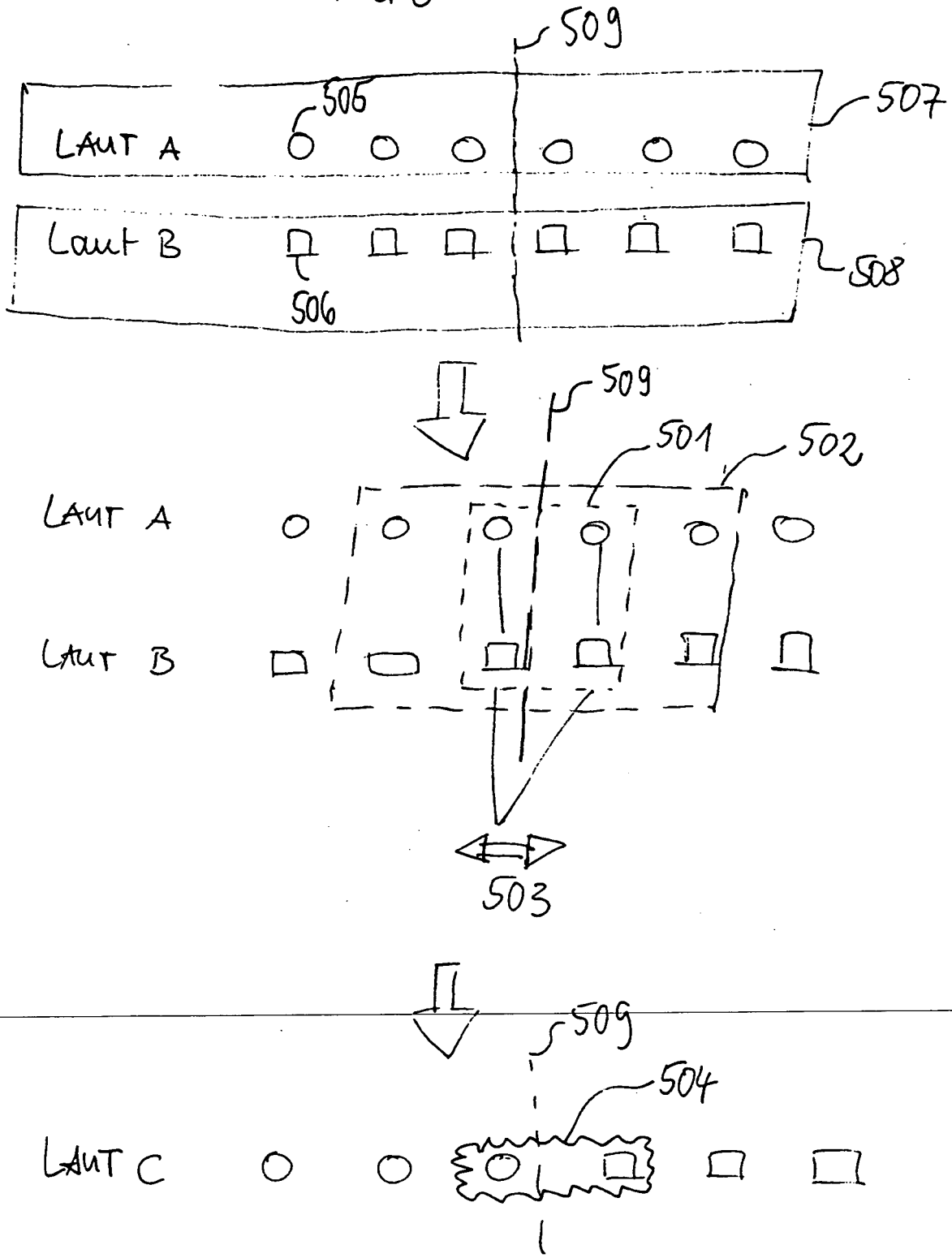


FIG 4

FIG 5


509
506
507
LAUT A
Laut B
506
508
509
501
502
LAUT A
LAUT B
503
509
504
LAUT C